Notes on Influence Maximization

immediate

1 Prior Work

Eshghi et al., "Efficient influence maximization under network uncertainty", 2019

Problem 1 (PV-IM)

Suppose there exists some underlying directed graph $G = (V, E)$ where $|V| = n, |E| = m$. If we are given

- *a visible induced subgraph of this network $G_v = (V_v, E_v)$ where $V_v \subseteq V$*
- *the structure of the unobserved part of the network is one of M different realizations $G_1, \ldots, G_M$ each with respective probability q_i such that $\sum_{i=1}^{M} q_i = 1$*
- *a diffusion model (either Independent Cascade or Weighted Cascade) with edge weight function $p\colon E \to [0, 1]$ and positive integer k*

find a seed set $S \subseteq V_v$ of size k that maximizes $\sigma(S)$. This is defined to be the a priori expected spread from S among all realizations (i.e. $\sigma(S) = \sum_{i=1}^{M} q_i \sigma_i(S)$ where $\sigma_i(S)$ is the expected spread of S in the graph $\mathcal{G}_i = G_v \cup G_i$).

The main contribution of this paper is to extend the idea of Reverse Reachable (RR) sets to this new setting by specifying how many RR sets need to be generated for each network realization $\mathcal{G}_i$. Let $x_i^j(S)$ be a binary variable denoting whether the jth RR set for $\mathcal{G}_i$ overlaps with any of the nodes in set S. Based on this we can define

$$F_i(S) = \frac{1}{\theta_i} \sum_{j=1}^{\theta_i} x_i^j(S)$$

to be the fraction of RR sets generated on $\mathcal{G}_i$ that S covers. This value is an unbiased estimator of $\sigma_i(S)/n$, which gives us that $F(S) = \sum_{i=1}^{M} q_i F_i(S)$ (i.e. the weighted average of the fraction of total RR sets that overlap with S) is an unbiased estimator of $\sigma(S)/n$.

Let S^* be an optimal seed set. Define p^* to be the expected influenced fraction of nodes in the network with S^* as a seed set (i.e. $p^* = \sigma(S^*)/n$). For each graph realization $\mathcal{G}_i$, define p_i^* similarly (i.e. $p_i^* = \sigma_i(S)/n$). Then $p^* = \sum_{i=1}^{M} q_i p_i^*$.

Lemma 1 ([1], Lemma 1)**.** *For $\delta_1 \in (0, 1)$ and $\varepsilon_1 > 0$, if*

$$\theta^* = \frac{\sum_{i=1}^{M} q_i^2 p_i^* (1 - p_i^*)}{\left(\sum_{i=1}^{M} q_i p_i^*\right)^2} \frac{\log(1/\delta_1)}{\varepsilon_1^2} = \frac{\sum_{i=1}^{M} q_i^2 p_i^* (1 - p_i^*)}{(p^*)^2} \frac{\log(1/\delta_1)}{\varepsilon_1^2}, \tag{1}$$

then for $\theta \geq \theta^*$, $nF(S^*) \geq (1 - \varepsilon_1)\sigma(S^*)$ with probability at least $1 - \delta_1$, where each realization

$\mathcal{G}_i$ *is sampled*

$$\theta_i^* = \theta^* \frac{q_i \sqrt{p_i^* (1 - p_i^*)}}{\sum_{j=1}^{M} q_j \sqrt{p_j^* \left(1 - p_j^*\right)}} \tag{2}$$

times for RR sets.

Lemma 1 implies that if a set S is constructed greedily from nodes covering these θ^* RR sets, then

$$nF(S) \geq \left(1 - \frac{1}{e}\right) nF(S^*) \geq \left(1 - \frac{1}{e}\right) (1 - \varepsilon_1)\, \sigma(S^*) = \left(1 - \frac{1}{e} - \varepsilon_1 \left(1 - \frac{1}{e}\right)\right) \sigma(S^*)$$

with probability $1 - \delta_1$. Intuitively, since $nF(S)$ is an indicator of $\sigma(S)$, this implies that S is likely to be large. However, the greedy algorithm can still construct a suboptimal seed set with some probability. To bound this possibility, we may need to generate more RR sets to ensure that the estimators of these suboptimal seed sets are close to their expected values.

Lemma 2. *For* $\delta_2 \in (0, 1)$, $\varepsilon > \varepsilon_1 \left(1 - \frac{1}{e}\right) > 0$, *if Lemma 1 holds and*

$$\theta' = \frac{2 \log\left(\frac{\binom{n}{k}}{\delta_2}\right) \left[(1 - \varepsilon_1)\left(1 - \frac{1}{e}\right) - \frac{2}{3}\varepsilon\right]}{\left(\varepsilon - \varepsilon_1 \left(1 - \frac{1}{e}\right)\right)^2 p^*}, \tag{3}$$

then for $\theta \geq \theta'$, $\sigma(S) \geq \left(1 - \frac{1}{e} - \varepsilon\right) \sigma(S^*)$ *with probability at least* $1 - \delta_2$, *where each realization* $\mathcal{G}_i$ *is sampled* $\theta_i' = q_i \theta'$ *for RR sets.*

Lemmas 1 and 2 then give the following main result.

Theorem 1. *For* $\varepsilon > \varepsilon_1 \left(1 - \frac{1}{e}\right) > 0$, $\delta_1, \delta_2 \in (0, 1)$ *and* $\delta = \delta_1 + \delta_2$, *if* $\mathcal{G}_i$ *is sampled* $\theta_i = \max\{\theta_i^*, \theta_i'\}$ *times for RR sets for all* i, *then the greedy algorithm returns a* $\left(1 - \frac{1}{e} - \varepsilon\right)$ *approximation with probability at least* $1 - \delta$.

2 Extensions

The actual construction of the RR sets do not necessarily depend on the diffusion model used. As long as there is a way to create the RR sets, then the same bounds on the number of RR sets needed for each $\mathcal{G}_i$ and the general algorithm of [1] can still be applied. [2] show that the triggering model still allows for the generation of RR sets. In general, if the diffusion model used does not admit a submodular influence spread function, then it is impossible to generate RR sets under that model.

In the original IMM paper [2], they provide a lower bound on the size of $\sigma(S^*)$ as that value is used in constructing the bounds on the number of RR sets needed. This paper does not provide that analysis.

References

[1] S. Eshghi, S. Maghsudi, V. Restocchi, S. Stein, and L. Tassiulas. "Efficient influence maximization under network uncertainty". In: *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE. 2019, pp. 365–371.

[2] Y. Tang, Y. Shi, and X. Xiao. "Influence maximization in near-linear time: A martingale approach". In: *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 2015, pp. 1539–1554.